Research Article Vol. 28, No. 13 / 22 June 2020 / Optics Express 19711
Optics EXPRESS

Human gesture recognition under degraded environments using 3D-integral imaging and deep learning

GOKUL KRISHNAN,[1] RAKESH JOSHI,[1] TIMOTHY O'CONNOR,[2] FILIBERTO PLA,[3] AND BAHRAM JAVIDI[1,*]

[1]*Electrical and Computer Engineering Department, University of Connecticut, 371 Fairfield Way Unit 4157, Storrs, Connecticut 06269, USA*
[2]*Biomedical Engineering Department, University of Connecticut, 260 Glenbrook Road, Unit 3247, Storrs, Connecticut 06269, USA*
[3]*Institute of New Imaging Technologies, Universitat Jaume I. Campus Riu Sec s/n, 12071 Castelló de la Plana, Spain*
[*]*bahram.javidi@uconn.edu*

Abstract: In this paper, we propose a spatio-temporal human gesture recognition algorithm under degraded conditions using three-dimensional integral imaging and deep learning. The proposed algorithm leverages the advantages of integral imaging with deep learning to provide an efficient human gesture recognition system under degraded environments such as occlusion and low illumination conditions. The 3D data captured using integral imaging serves as the input to a convolutional neural network (CNN). The spatial features extracted by the convolutional and pooling layers of the neural network are fed into a bi-directional long short-term memory (BiLSTM) network. The BiLSTM network is designed to capture the temporal variation in the input data. We have compared the proposed approach with conventional 2D imaging and with the previously reported approaches using spatio-temporal interest points with support vector machines (STIP-SVMs) and distortion invariant non-linear correlation-based filters. Our experimental results suggest that the proposed approach is promising, especially in degraded environments. Using the proposed approach, we find a substantial improvement over previously published methods and find 3D integral imaging to provide superior performance over the conventional 2D imaging system. To the best of our knowledge, this is the first report that examines deep learning algorithms based on 3D integral imaging for human activity recognition in degraded environments.

1. Introduction

Human gesture recognition involves deriving meaningful inference from human motions and has a wide range of applications in human-computer interaction, patient monitoring, surveillance, robotics, sign language recognition, etc. [1]. In recent years, human gesture and action recognition have attracted wide interest among the computer science and computer vision communities. Numerous approaches have been proposed for gesture recognition, including the use of mathematical models such as Hidden Markov Models (HMM) [2], spatio-temporal interest points-based detectors (STIPs) [3], correlation filter-based approaches [4], etc. Recently, deep learning-based models for gesture recognition have gained wide acceptance due to their generalization capabilities and high accuracy in detecting and classifying gestures [5,6]. While these methods have been shown to work well for clean datasets, gesture recognition under degraded conditions remains a challenge, especially in cases where gestures are partially occluded, or in low illumination conditions. In degraded environments, the features of the gestures may not be

#396339 https://doi.org/10.1364/OE.396339
Journal © 2020 Received 28 Apr 2020; revised 14 Jun 2020; accepted 14 Jun 2020; published 19 Jun 2020

Optics EXPRESS

fully recorded during the camera pickup process, which can make gesture recognition under such conditions more challenging.

Most of the state-of-the-art vision-based gesture recognition methodologies consider non-occluded cases and typically only contain a single gesture present in the scene [1]. However, in real-world scenarios, multiple gestures may be present, and in many cases, the gestures may be partially occluded, and the illumination conditions may not be perfect. In multi-gesture scenarios, hand-to-hand occlusions may occur, which can make gesture detection difficult. The three-dimensional (3D) reconstruction algorithm based on integral imaging provides an efficient way of removing occlusions and detecting gestures [7,8]. In addition, our experiments suggest that the variable focusing and depth-sectioning capabilities of 3D integral imaging aid in the gesture detection capabilities when multiple gestures are present. Therefore, in this paper, we leverage the advantages of passive 3D integral imaging with deep learning to propose an efficient human multi-gesture recognition system under degraded conditions. The degraded conditions considered in this paper include partial occlusions in low illumination conditions.

Convolutional neural networks (CNNs) are a class of deep neural networks, widely used for tasks such as object recognition and image classification [5,9]. As the name suggests, CNNs consist of stacked convolutional and pooling layers followed by one or more fully connected layers and a classification layer. The convolutional layers use a series of convolutional kernels for feature learning. The output of the convolutional layers will be feature vectors, which represent the spatial features representing an image. The pooling layer combines semantically similar features into a single feature [9]. In the case of gesture recognition, the temporal dependency of feature vectors extracted from adjacent frames are also important, which CNN alone cannot capture. Therefore, we have used a cascaded network comprising of CNN and bi-directional long short-term memory (BiLSTM) network. The convolutional layers of the CNN produce the feature vectors, which are used by the BiLSTM network for gesture classification. Furthermore, we compare the results of the proposed deep learning-based method with respect to the previously reported 3D integral imaging-based spatio-temporal interest point with support vector machines (STIP-SVM) [8] and distortion invariant non-linear correlation filter-based approaches [7].

This paper is organized into four different sections: Section 1 provides the introduction and a brief review of gesture recognition and the various approaches by which it may be achieved. Section 2 discusses the proposed approach and its details. Section 3 deals with the experimental results, including the performance of the proposed system, and comparison with previously reported methodologies. Finally, section 4 provides the conclusions of the paper.

2. Methodology

2.1. Computational volumetric reconstruction using integral imaging (InIm)

Integral imaging is a passive three-dimensional (3D) imaging technique that captures both the intensity and directional information of a scene using an array of cameras, lenslet array, or a moving camera system [10–16]. Originally proposed by Lippmann, the integral imaging-based techniques have proved to be useful for human action and gesture recognition [7,8]. In the original work, in which the method was defined as integral photography, Lippmann used a microlens array (MLA) in front of photographic film in order to capture the multiple 2D images, each having a different perspective of the scene [16,17]. The individual 2D images are usually called elemental images (EIs). Following the advancements in digital sensors to replace photographic film, this technique is now generally referred to as integral imaging. As in the original conception, integral imaging may still be performed using a lenslet array in front of an imaging sensor. However, another interesting approach for implementing the concept of integral imaging to capture the 3D information is by using an array of digital cameras [18]. The advantage of this approach is that the captured 3D scenes can have higher parallax and better resolution at a longer distance. Another possibility is using a single camera on a moving translation stage. This has been named as the

Optics EXPRESS

synthetic-aperture integral imaging [17], but this method is not suitable for dynamic scenes. By capturing the scene from multiple viewing perspectives, integral imaging helps in reducing noise due to partial occlusions [19], low-light illumination [20,21], and scattering medium [22–24], etc. The computational reconstruction algorithm is an inverse mapping procedure, which integrates the elemental images by back-projecting the elemental images through a virtual pinhole array into the object space at the desired reconstruction depth [10]. It has been shown that this process is optimal in a maximum likelihood sense [17,25]. The integral imaging pickup process and the computational reconstruction process are shown in Fig. 1(a) and Fig. 1(b), respectively. The spatio-temporal volume video data is reconstructed as follows [7]:

$$r(x,y,z;t) = \frac{1}{O(x,y;t)} \sum_{i=0}^{K-1} \sum_{j=0}^{L-1} EI_{i,j}\left(x - i\frac{r_x \times p_x}{M \times d_x},\ y\ -j\frac{r_y \times p_y}{M \times d_y}; t\right) \tag{1}$$

where, $r(x,y,z;t)$ is the integral imaging reconstructed video. The reconstructed video is obtained by shifting and overlapping $K \times L$ elemental images at reconstruction depth. The x, y represents the pixel indices and t is the frame index. Here, $EI_{i,j}$ represents the i,j^{th} elemental image. The r_x, r_y and d_x, d_y in Eq. (1) represent the resolution and physical size of the image sensor, respectively. The p_x, and p_y indicate the pitch of adjacent image sensors on the camera array in the x, and y directions, respectively. The $O(x,y;t)$ matrix contains information regarding the number of overlapping pixels, and the magnification factor is $M = \frac{z}{f}$, where f represents the focal length.

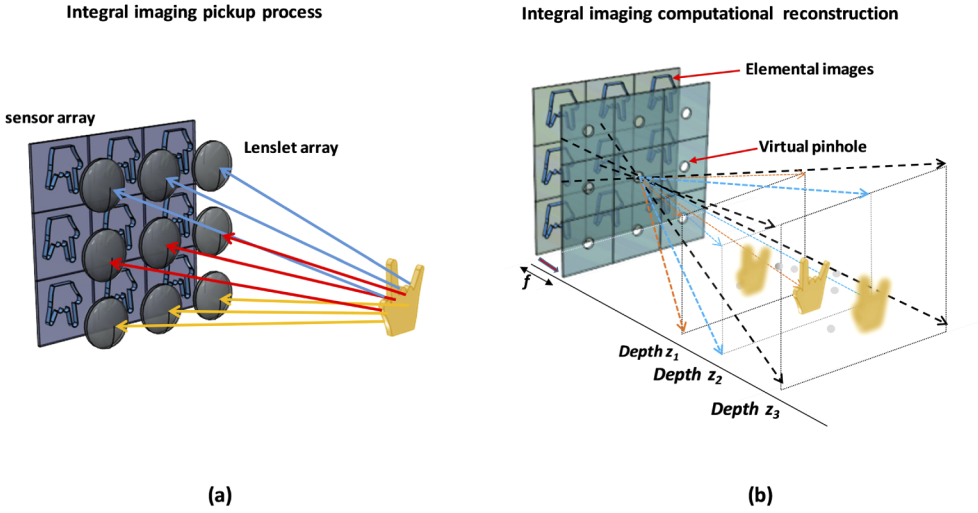

Fig. 1. (a) Integral imaging pickup process, (b) computational volumetric reconstruction process using integral imaging.

2.2. *Video encoding using convolutional neural network (CNN)*

The 3D reconstructed video obtained using Eq. (1) is encoded using a CNN. The CNN consists of a series of convolutional and pooling layers, which spatially filter the input data to yield relevant features. The output of the last pooling layer produces the feature vector representing the spatial information of the video. This feature vector can be used for high-level interpretation tasks such as object detection and classification. In order to mitigate the effect of lesser training data, we have used the deep GoogLeNet network pretrained on the well-known ImageNet [26] dataset for feature extraction. The GoogLeNet uses an inception module. The inception module architecture is similar to that proposed by Serre et al. [27], which uses a series of Gabor filters to handle

Optics EXPRESS

multiple scales. The filters in the inception module are learned, and the inception layers are repeated multiple times in order to get a deep neural network model [28]. The inception module consists of multiple filters in parallel, 1×1, 3×3, 5×5 convolution, and max-pooling operations [28]. The resulting filter outputs are concatenated. The input size of the network is 224×224 and takes the RGB color channels with the mean subtracted. The CNN architecture we used is 22 layers deep, excluding the pooling layers. As we go deeper, the feature representations become more abstract and informative. We have used a total of 9 inception modules and global average pooling in order to generate the feature estimates. A series of max-pooling layers have been stacked in between the inception modules in order to reduce the dimensionality of the features learned at each stage. The convolution layers, including those present in the inception module use the rectified linear unit (ReLU) as the activation function [28].

Assuming the 3D reconstructed video consists of N frames, for each frame I_n, $= \{1, 2, \ldots, N\}$, the convolutional and pooling layers of the CNN produces a K dimensional feature vector $\mathbf{x}_i \in \mathbb{R}^K$, $i = 1, 2, \ldots, N$. Therefore, when also considering the temporal dimension of the data, the input can be represented by a feature matrix $\mathbf{X} \in \mathbb{R}^{K \times N}$, i.e.

$$\mathbf{X} = \mathbf{x}_1 \odot \mathbf{x}_2 \odot \mathbf{x}_3 \odot \ldots \odot \mathbf{x}_N \tag{2}$$

where, $\odot$ represents the concatenation operator. Thus, for each input video, the CNN produces the corresponding feature matrix $\mathbf{X}_j \in \mathbb{R}^{K \times N}$, $j = \{1, 2, \ldots, P\}$, where P represents the number of input videos. The rows and columns of $\mathbf{X}_j \in \mathbb{R}^{K \times N}$ encode the spatial and temporal information of each input video, respectively.

2.3. LSTM network

For applications involving sequential inputs, such as in gesture recognition, the recurrent neural networks (RNNs) have been shown to be useful for capturing the temporal dependency of the data. Recurrent neural networks (RNNs) are a very powerful dynamical system, but training them has been a difficult task due to the problem of vanishing and exploding gradients [9]. In order to deal with the problems present in conventional RNNs, Hochreiter and Schmidhuber [29] proposed long short-term memory networks (LSTMs). These networks have improved the "memory" of the cell by introducing a "gate" into it. Consider an input sequence of T time steps, $\mathbf{x} = [x_1, x_2, x_3, \ldots, x_T]$ which is fed to a recurrent neural network, a standard RNN computes the hidden vector $\mathbf{h} = [h_1, h_2, h_3, \ldots, h_T]$ and the output vector $\mathbf{h} = [h_1, h_2, h_3, \ldots, h_T]$ as follows [30]:

$$h_t = \sigma(\boldsymbol{W}_{ih} x_t + \boldsymbol{W}_{hh} h_{t-1} + \boldsymbol{b}_h) \tag{3}$$

and

$$y_t = \boldsymbol{W}_{ho} h_t + \boldsymbol{b}_o \tag{4}$$

for time instants $t = (1, 2, 3, \ldots, T)$. In the above equations, $\sigma(x) = 1/(1 + e^{-x})$ represents the element-wise logistic sigmoid function. Here, $\mathbf{W}_{kk}$ and $\mathbf{b}_k$, $k \in \{o, c, f, i, x, h\}$ represent the corresponding weight matrices and bias terms of the network, respectively. Unlike standard RNNs, LSTMs [29] use memory cells to extract the long-term temporal relationships hidden inside the sequential input. The memory cells have a multiplicatively gated self-connection with unity weight, which copies its state and accumulates the external signal. This multiplicative gating is controlled by another unit that learns to decide when to clear the memory [9]. A standard LSTM network computes the hidden vector using the following relationships [29,30],

$$i_t = \sigma(\boldsymbol{W}_{xi} x_t + \boldsymbol{W}_{hi} h_{t-1} + \boldsymbol{b}_i) \tag{5}$$

$$f_t = \sigma(\boldsymbol{W}_{xf} x_t + \boldsymbol{W}_{hf} h_{t-1} + \boldsymbol{b}_f) \tag{6}$$

$$c_t = f_t c_{t-1} + i_t \tanh(\boldsymbol{W}_{xc} x_t + \boldsymbol{W}_{hc} h_{t-1} + \boldsymbol{b}_c) \tag{7}$$

Optics EXPRESS

$$o_t = \sigma(\boldsymbol{W}_{xo}x_t + \boldsymbol{W}_{ho}h_{t-1} + \boldsymbol{b}_o) \tag{8}$$

$$h_t = o_t \tanh(c_t) \tag{9}$$

where, i, f, o, are respectively the input, forget and output gates, and c is the cell state vector.

The bi-directional long short-term memory (BiLSTM) network [31] is a variant of long short-term memory (LSTM) network, where two separate networks are used to extract all the available input information. It has been previously reported that the BiLSTM networks outperform unidirectional LSTM networks [32]. In a BiLSTM network, one network will be responsible for learning in the forward time direction, which outputs $\mathbf{h}_{forward}$ while the other will be for learning in reverse time direction outputs $\mathbf{h}_{reverse}$. The output of these two layers will be merged, i.e., $\mathbf{h} = f(\mathbf{h}_{forward}, \mathbf{h}_{reverse})$. We have chosen the merging operator f to be a concatenation operator; therefore, $\mathbf{h} = [\mathbf{h}_{forward}, \mathbf{h}_{reverse}]$. The structure of an LSTM cell and the BiLSTM architecture has been shown in Fig. 2(a) and Fig. 2(b), respectively. Finally, the merged output is fed to a fully connected layer and a Softmax layer for classification purposes. We used a BiLSTM layer with 100 hidden units to learn the temporal dependency of the feature vectors. The recurrent weights of the network are randomly initialized from a unit normal distribution. We have used hyperbolic tangent (*tanh*) and sigmoid function for the state activation function and the gate activation function, respectively. The forward and backward layers of BiLSTM network do not interact with each other. This allows us to use the standard training algorithms used for RNNs to train our network. The network was optimized using Adam optimizer with a gradient decay factor of 0.9, and a learning rate of 10^{-4}. To mitigate overfitting, we have used the dropout

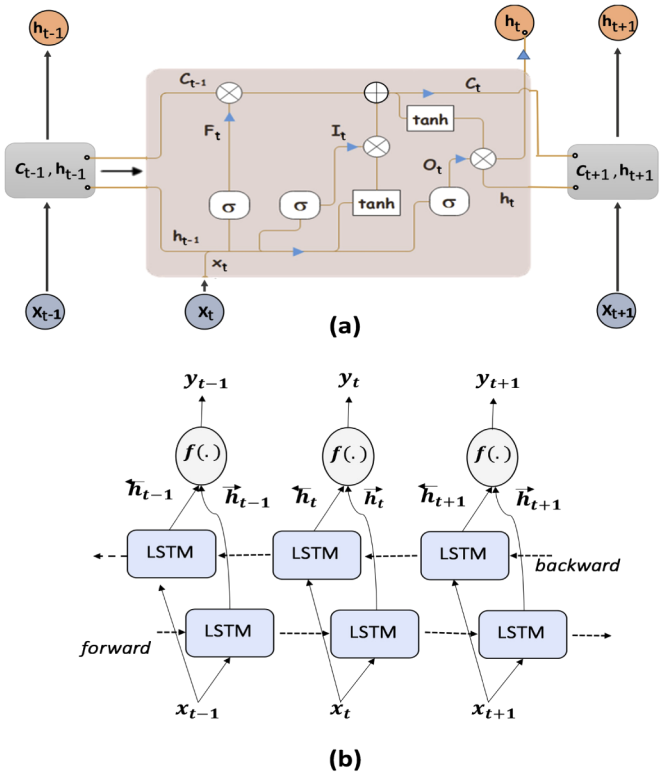

Fig. 2. (a) Structure of a long short-term memory (LSTM) cell (b) a bi-directional long short-term memory (BiLSTM) network architecture.

Optics EXPRESS

layer with a dropout rate of 0.5. The network was trained for 25 training epochs with a batch size of 20 for each training iteration.

The block diagram of the proposed system is shown in Fig. 3. The human gesture data has been recorded using the integral imaging technique. The 3D computational reconstruction algorithm has been used to reconstruct the data at the correct depth. The reconstructed data is fed into a convolutional neural network for feature extraction. The convolution and pooling layers of the network extract a set of spatial feature vectors that are input into a Bi-directional LSTM network (BiLSTM) that is designed to capture the temporal variation between the feature vectors. Finally, the Softmax and classification layers categorize different gestures.

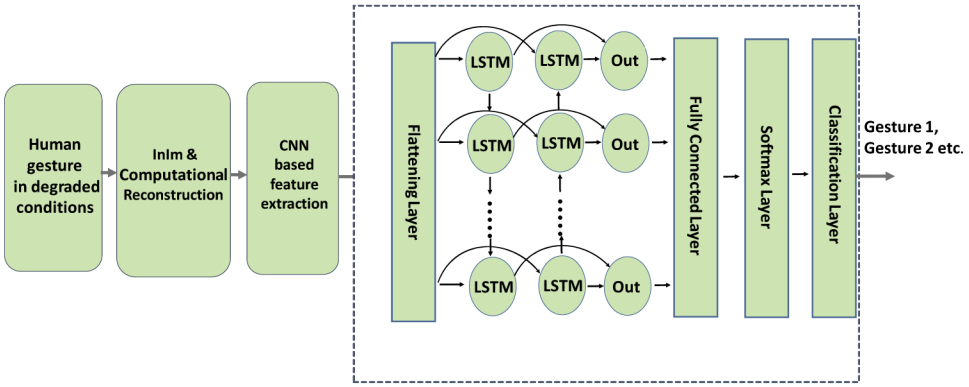

Fig. 3. Block diagram for Integral imaging-based gesture recognition using proposed deep learning architecture. InIm: Integral imaging, CNN: convolutional neural network, LSTM: Long Short-Term Memory Network.

3. Experimental results and discussions

In this section, we discuss the performance of the proposed approach on experimental data. To that end, a 3×3 camera array was used for integral imaging, as shown in Fig. 4(a). In our experiments, we used Mako G192C machine vision cameras. All the cameras have identical intrinsic parameters. The pitch of the camera array was designed to be 80 mm in both x and y directions. The pixel size is 4.5 µm × 4.5 µm, the resolution of each camera is 1200 × 1600, and the focal length was set to be 15 mm. The quantum efficiency of the camera is 0.44 at a wavelength of 525 nm, and the sensor read noise is 20.47 electrons rms/pixel. All the cameras in the array are synchronized to record the data with a frame rate of 10 frames per second (fps). The data is recorded with camera lens F-number of 1.8 and an exposure time of 30 milliseconds (ms). In our experiments, we considered two classes of gestures in order to demonstrate the potential benefits of the proposed approach. The gesture motions are depicted in Fig. 4(b). The data were collected from 4 participants with 5 different backgrounds, as shown in Fig. 4(c). The gestures are captured at a distance of about 3 meters away along the axial direction from the camera array. The participants were asked to repeat each gesture twice to capture both fast and slow variations of the same gesture. The low illumination effects considered in the experiments have been simulated by computational models applied to the experimentally captured elemental images in order to generate a large number of low light data for testing.

We have conducted three different experiments and collected data under three different conditions: 1) single gesture without any occlusion, 2) single gesture with partial occlusion, 3) multiple gestures in the scene at different depths without added occlusion, plus low light effects. For all the experiments, we assume the depth of gesture of interest is known a priori. This assumption is for convenience, and it is not necessary as integral imaging can reconstruct

Optics EXPRESS

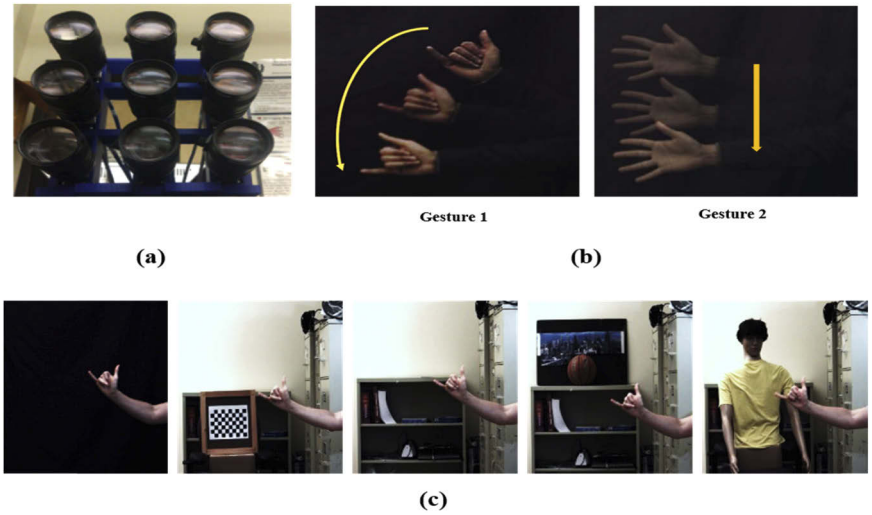

Fig. 4. Figures showing (a) 3 × 3 camera array for integral imaging capture stage used for our experiments, (b) depicts the two different gesture motions considered in this paper, (c) shows a single gesture with different scene backgrounds used for our experiments.

the in-focus object of interest in the 3D scene. In the first experiment, we collected data without occlusion, as shown in Fig. 5(a) and in the second experiment, we recorded data with partial occlusion, as shown in Fig. 5(b). The integral imaging-based computational reconstruction algorithm, as outlined in section 2.1, has been used to reconstruct the 3D data at the correct depth, where the gesture of interest is located. As shown in Fig. 5(c), the integral imaging technique helps us to remove the effect of partial occlusion and thereby improving the visibility and thus enhancing the classification performance.

To study the performance of the proposed system in low illumination conditions, we have computationally simulated the low illumination conditions using the degradation model $I_{degraded} = \alpha \times I_{original} + n$, where α is the attenuation factor, and n represents the camera noise, modeled as Gaussian noise with mean μ and variance σ^2. For our analysis, we considered the attenuation factor $\alpha = 0.2$. The Gaussian noise with mean μ equal to zero and variance equal to 0.05 has been added after attenuating the videos. In 2D imaging, a single frame representing the low illumination condition for gesture 1 and gesture 2 has been shown in Fig. 5(d). The signal-to-noise ratio (*SNR*) for the corresponding simulated low illumination images is calculated using [21,33], $SNR = (\mu_s - \mu_n)/\sqrt{\sigma_s^2 + \sigma_n^2}$, where μ_s, μ_n correspond to the means of the signal (the object of interest) area and the background area, respectively, and σ_s^2 , σ_n^2 denote their corresponding variances. Thus, the *SNR* calculated for the degraded image shown in Fig. 5(d) is 0.0438 and 0.0643 for gesture 1 and gesture 2, respectively. The number of photons per pixel ($N_{photons}$) for the degraded image can be calculated as [7,21]:

$$N_{photons} = SNR * n_r/QE \tag{10}$$

where, n_r represents the camera read noise, and QE represents the quantum efficiency of the camera. Using Eq. (10), the estimated photons per pixel for the degraded 2D frame is 2.0376 and 2.99 photons per pixel for gesture 1 and gesture 2, respectively. The integral imaging computational reconstruction algorithm is used to reconstruct the degraded image at the depth of the object of interest, as shown in Fig. 5(e). The *SNR* of the 3D reconstructed frame for gesture 1 and gesture 2 are 0.1404 and 0.1699, respectively.

Optics EXPRESS

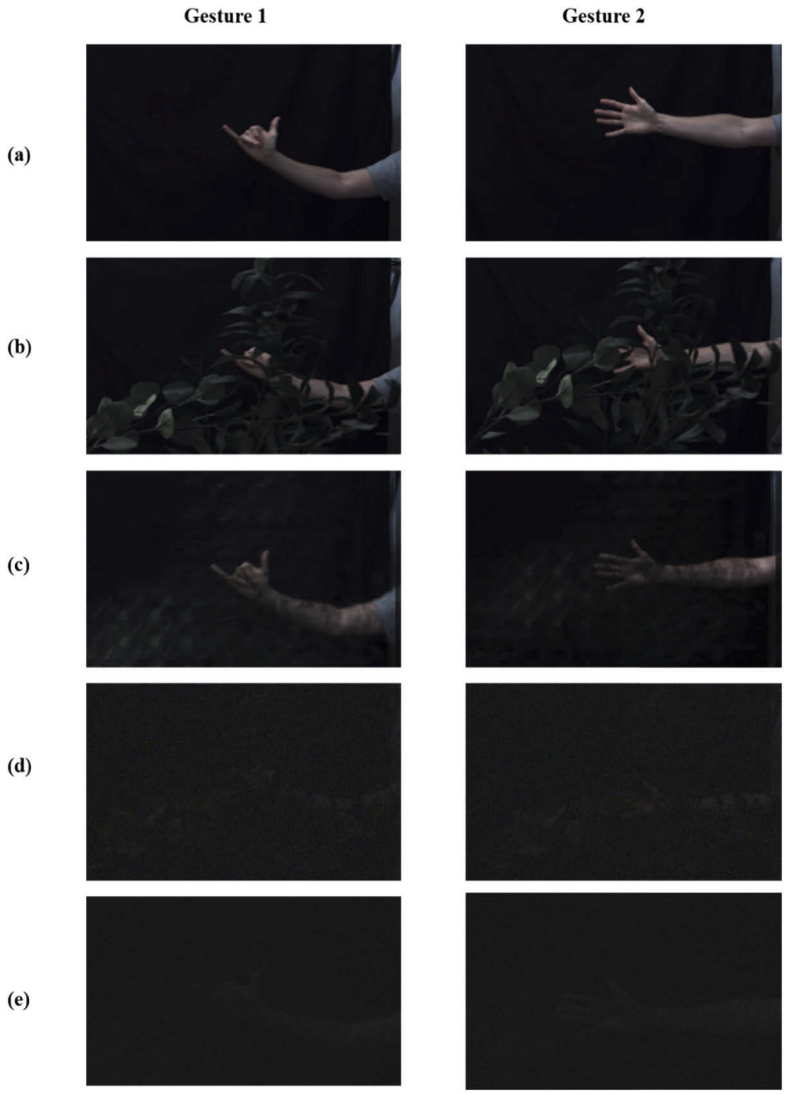

Fig. 5. Images of gesture 1 and gesture 2 (a) 2D imaging without occlusion (b) 2D imaging with occlusion (c) reconstructed image using integral imaging with occlusion (d) 2D imaging with occlusion under simulated low illumination condition (Gesture 1: SNR = 0.0438, Gesture 2: SNR = 0.0643) (e) 3D Integral imaging reconstruction with occlusion under simulated low illumination (Gesture 1: SNR = 0.1404, Gesture 2: SNR = 0.1699).

Optics EXPRESS

In the next experiment, we considered a multiple-gesture scenario. In this case, we considered two gestures present in the scene. One of the gestures is considered as the gesture of interest (true class), and the other is considered a random false class gesture happening elsewhere in the scene. The two gestures are assumed to be happening at different depths. Since a random gesture can happen anywhere in the scene, for our experiments, we have chosen three different locations for the background gesture, as shown in Fig. 6.

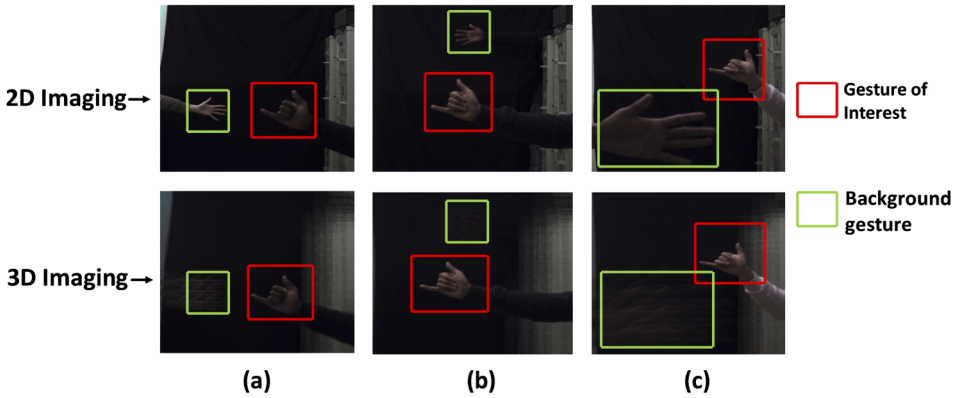

Fig. 6. Scene for multiple gestures experiments. Three different locations for the background gesture (false class) considered in our experiments. (a) behind and to the left of the true class gesture of interest, (b) behind and to the right of the true class gesture of interest, and (c) in front and to the left of the true class gesture of interest.

To study the effect of low illumination degradation, we have computationally simulated the low light degradation condition for the multi-gesture scenarios. The degradation has been simulated using the same degradation model as described in the above paragraph, by attenuating the recorded video and adding additive Gaussian noise. Figure 7 depicts one such scenario where both gestures are clearly visible in the case of conventional 2D imaging with normal illumination, as shown in Fig. 7(a), whereas the 3D integral imaging in-focus reconstruction allows us to extract the gesture of interest from the other out-of-focus objects in the scene as shown in Fig. 7(b). The low light 2D frame is shown in Fig. 7(c) with *SNR* equal to 0.1092 and 0.1413 for gesture 1 and gesture 2, respectively. Using Eq. (10), the photons per pixel ($N_{photons}$) has been calculated and equals to 5.0802 and 6.5735 photons per pixel, for gesture 1 and gesture 2, respectively. As shown in Fig. 7(d), the integral imaging reconstruction algorithm reduces noise and improves the visibility in low illumination. The *SNR* of the 3D reconstructed frames for gesture 1 and gesture 2 are 0.2957 and 0.3875, respectively.

In total, 240 gesture videos were recorded, 80 videos (2 gestures x 4 participants x 5 backgrounds x 2 repetitions) corresponding to each of the three experiments as following: (1) no occlusion, (2) with partial occlusion, and (3) multi-gesture. Out of the 80 videos, 40 videos belong to the first gesture and 40 videos belong to the second gesture. The data collected without occlusion has been used for training the system since it is assumed that only *clean* datasets may be available for training the network, as one should not expect to know the occlusion a priori in a real-world scenario. The degraded datasets, such as those degraded conditions considered in this report, appear unpredictably over time and may be unknown to the system. To improve the performance and the generalization capabilities of the neural network model, we have used data augmentation to increase the size of the training dataset. We have used six different data augmentation techniques, including affine transformation, flipping, resizing, inversion, blurring, and noising. Thus, after data augmentation, we generated a total of 560 videos for both gesture classes, which we used for training the model. We have tested our model using the data with

Optics EXPRESS

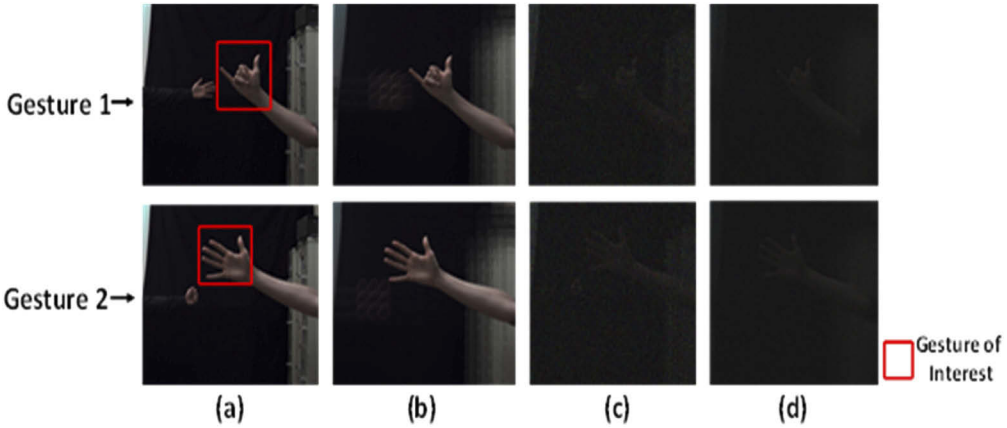

Fig. 7. Visualization of the multi-gesture scenario. (a) 2D elemental image, (b) 3D reconstructed image using integral imaging, (c) 2D elemental image in low illumination (Gesture 1: SNR = 0.1092, Gesture 2: SNR = 0.1413), (d) 3D reconstructed image in low illumination using integral imaging (Gesture 1: SNR = 0.2957, Gesture 2: SNR = 0.3875).

occlusion, the occlusion data with simulated low illumination, the multi-gesture data, and the multi-gesture data with simulated low illumination. We have obtained the receiver operating characteristics (ROC) curves for comparing the proposed technique with 1) a conventional 2D imaging-based deep learning technique, 2) the previously reported STIP-SVM approach [8], and 3) the distortion invariant non-linear correlation-based approach (using k = 0.3) [7,34,35]. For comparing the performance of different classifiers using ROC analysis, we use the area under the ROC curve (AUC), which is widely used for assessing the overall performance of classifiers based on their ROC curves [36,37]. For data with occlusion, we achieved an area under the curve (AUC) of 0.958 using the proposed 3D integral imaging-based CNN-bi-directional long short-term memory (CNN-BiLSTM) technique. For 2D imaging-based CNN-BiLSTM and 3D integral imaging-based STIP-SVM approaches, we obtained AUC values of 0.696 and 0.572, respectively. In the case of occlusion in simulated low illumination conditions, the proposed 3D technique yields an AUC of 0.941 while the 2D imaging-based CNN-BiLSTM and 3D integral imaging-based STIP-SVM approach each yielded an AUC of 0.574 and 0.476, respectively. For the cases of environmental degradation, such as occlusion and occlusion with low light illumination, the distortion invariant non-linear correlation approach produces an AUC of 0.471 and 0.477, respectively. These eight ROC curves for the scene with occlusion data are illustrated for comparison in Fig. 8.

The effect of low illumination on the performance of different methodologies has been studied based on the percent reduction in area under the curve (AUC). In our experiments, the introduction of low illumination reduces the AUC by 1.77%. For the proposed 3D integral imaging-based CNN approach, while for 2D imaging and STIP-based approaches, the AUC decreases by 17.53% and 16.78%, respectively. In the case of distortion invariant non-linear correlation approach, the percentage change in AUC is 1.27%.

The ROC analysis for the multi-gesture scenario is presented in Fig. 9. We have presented eight ROC curves for comparison. For the multi-gesture case, the proposed 3D CNN-BiLSTM approach achieves an AUC of 0.998 and in the multi-gesture with the low-illumination scenario, the proposed approach yields an AUC of 0.964. The 2D imaging-based CNN-BiLSTM and 3D integral imaging-based STIP-SVM approaches give AUC values of 0.958, and 0.426 respectively, for the multi-gesture scenario, and 0.705, and 0.519, respectively, for the multi-gesture data with simulated low-illumination conditions. The non-linear correlation approach produces an

Optics EXPRESS

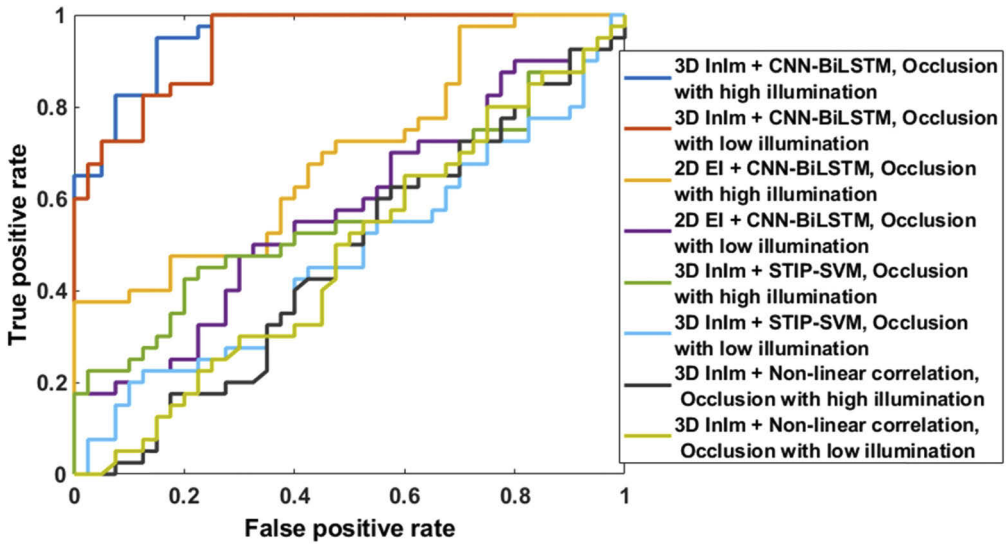

Fig. 8. ROC curves for comparison of different gesture classification methodologies under partial occlusion. (a) 3D integral imaging (InIm)-based CNN-BiLSTM network in high illumination condition (3D InIm + CNN-BiLSTM, dark blue line), (b) 3D integral imaging-based CNN-BiLSTM with low illumination (3D InIm + CNN-BiLSTM, red line), (c) 2D elemental imaging (EI)-based CNN-BiLSTM in high illumination condition (2D EI + CNN-BiLSTM, yellow line) (d) 2D elemental imaging (EI)-based CNN-BiLSTM with low illumination (2D EI + CNN-BiLSTM, violet line), (e) 3D integral imaging (InIm)-based spatio-temporal interest points (STIP) - SVM classifier in high illumination condition (3D InIm + STIP-SVM, green line), (f) 3D integral imaging (InIm)-based spatio-temporal interest points (STIP) - SVM classifier with low illumination (3D InIm + STIP-SVM, light blue line), (g) 3D integral imaging (InIm)-based distortion invariant non-linear correlation (k = 0.3) approach in high illumination (3D InIm + Non-linear correlation, black line), (h) 3D integral imaging (InIm)-based distortion invariant non-linear correlation (k = 0.3) approach with low illumination (3D InIm + Non-linear correlation, light green line).

AUC of 0.467 and 0.444 for multi-gesture scenarios in high illumination and low illumination, respectively. As in the case of occlusion, the effect of degradation on the multi-gestures case has been analyzed by the percentage reduction in AUC due to degradation. The percentage reduction in AUC for the proposed approach is 3.41%, while the percentage change for 2D imaging and STIP-based SVM technique is 26.41% and 21.83%, respectively. For the case of distortion invariant non-linear correlation approach, the percentage change in AUC is 4.93%.

In the multi-gesture scenario, we have only considered experimentally the cases wherein the false gesture takes place either in front of or behind the gesture of interest. However, the approach may also be effective when both gestures occur at the same plane, especially in degraded conditions. The integral imaging reconstruction acts as a depth-based filter and removes noise from the depths outside the depth of reconstruction. This depth-sectioning may improve performance over conventional 2D imaging strategies regardless of the depth of the false gesture. In particular, under degraded environments such as low-illumination or partial occlusion, the 3D integral imaging helps to mitigate the effects of degradation. Thus, the proposed approach provides better performance as compared to the 2D imaging-based CNN-BiLSTM and the other approaches used for comparison.

The STIPs-based feature extraction is defined based on the variations of local gradients in spatial and temporal dimensions. Thus, the STIPs are suitable for detecting moving objects in

Optics EXPRESS

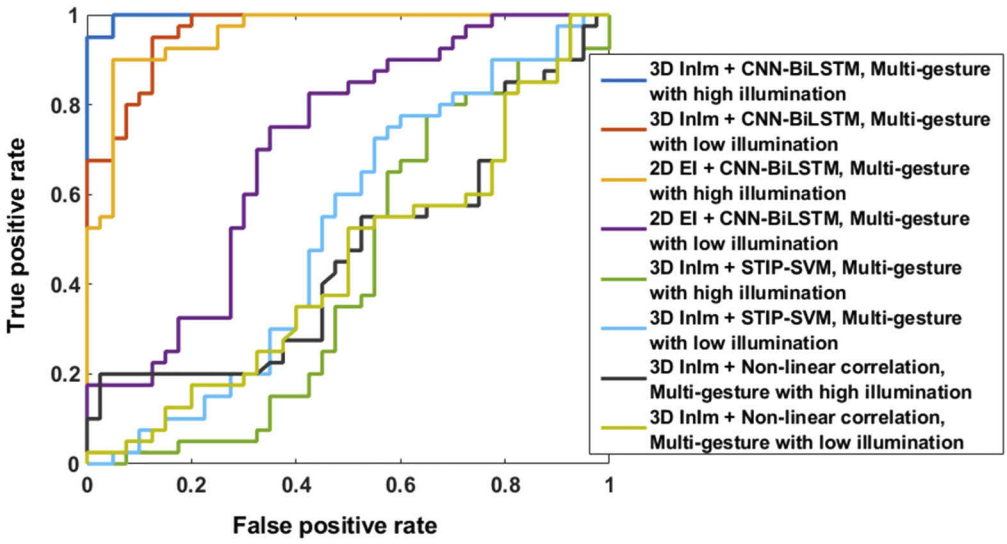

Fig. 9. ROC curves for comparison of different gesture classification methodologies in multi-gestures scenario. (a) 3D integral imaging (InIm)-based CNN-BiLSTM network in high illumination condition (3D InIm + CNN-BiLSTM, dark blue line), (b) 3D integral imaging-based CNN-BiLSTM network with low illumination (3D InIm + CNN-BiLSTM, red line), (c) 2D elemental imaging (EI)-based CNN-BiLSTM network in high illumination condition (2D EI + CNN-BiLSTM, yellow line), (d) 2D elemental imaging (EI) based CNN-BiLSTM with low illumination (2D EI + CNN-BiLSTM, violet line), (e) 3D integral imaging (InIm)-based spatio-temporal interest points (STIP) - SVM classifier in high illumination condition, (3D InIm + STIP-SVM, green line), (f) 3D integral imaging (InIm)-based spatio-temporal interest points (STIP) - SVM classifier with low illumination (3D InIm + STIP-SVM, light blue line), (g) 3D integral imaging (InIm)-based distortion invariant non-linear correlation (k = 0.3) approach in high illumination (3D InIm + Non-linear correlation, black line), (h) 3D integral imaging (InIm)-based distortion invariant non-linear correlation (k = 0.3) approach with low illumination (3D InIm + Non-linear correlation, light green line).

a video, including for human gesture recognition tasks [8]. In our experiments, this approach achieves better performance for the single gesture case, as shown in Fig. 8 compared to the multi-gesture scenario, as shown in Fig. 9. In low light illumination conditions, the number of STIPs detected will be greatly reduced. Thus, the performance decreases in low illumination for the single gesture scenario, as shown in Fig. 8. Under the multi-gesture scenario, even though the integral imaging helps to mitigate the effects of the false gesture to a certain extent, there will still be local gradient variations due to the false gesture, especially when it is moving. The STIPs detected from the false gesture may still be prominent, especially in cases where the false gesture is present in front of the gesture of interest. Thus, the false classification may increase when the STIPs are detected from false gesture, in addition to the STIPs detected from the gesture of interest. In low light illumination, the number of STIPs detected will be much less, so the recognition performance is similar to that of a random classifier. As shown in Table 1, the STIP-based SVM recognition methods did not achieve an accuracy above 55.00% or an AUC above 0.572 in any case. These results suggest that the STIP-based SVM recognition method may not be an effective gesture recognition strategy for the experimental conditions considered in this work.

From Fig. 8 and Fig. 9, our experimental comparison shows that while the previous works demonstrated that the correlation-based approach outperformed the STIP-based recognition

Optics EXPRESS

Table 1. Comparison of gesture recognition performance under various experimental conditions.[a]

Method	Condition	Accuracy	AUC	F1 score	MCC
Proposed 3D integral imaging-based CNN – Bidirectional Long Short-Term Memory approach	Occlusion	86.25%	0.958	0.874	0.736
	Occlusion with low illumination	83.75%	0.941	0.860	0.714
	Multi-gesture scenario	96.25%	0.998	0.962	0.925
	Multi-gesture with low illumination	83.75%	0.964	0.817	0.693
2D imaging-based CNN – Bidirectional Long Short-Term Memory approach	Occlusion	58.75%	0.696	0.535	0.180
	Occlusion with low illumination	47.50%	0.574	0.644	0.160
	Multi-gesture scenario	76.25%	0.958	0.708	0.566
	Multi-gesture with low illumination	62.50%	0.705	0.722	0.350
3D integral imaging with STIP- SVM approach [8].	Occlusion	55.00%	0.572	0.539	0.100
	Occlusion with low illumination	48.75%	0.476	0.328	0.024
	Multi-gesture scenario	53.75%	0.426	0.584	0.077
	Multi-gesture with low illumination	52.50%	0.519	0.513	0.050
3D integral imaging with distortion invariant non-linear correlation filter [7].	Occlusion	50.00%	0.471	0.524	-0.001
	Occlusion with low light illumination	50.00%	0.477	0.512	0.000
	Multi-gesture scenario	53.75%	0.467	0.546	-0.009
	Multi-gesture with low light illumination	46.25%	0.444	0.460	0.006

[a]Abbreviations: CNN – convolutional neural network, STIP – spatio-temporal interest points, SVM – support vector machine, AUC – Area under the ROC curve, MCC- Mathew's correlation coefficient.

method, under more challenging experimental conditions including occlusion and low light illumination, both STIP-based and correlation-based gesture recognition strategies may be less effective than the proposed CNN-BiLSTM approach. The correlation-based classification is particularly useful for narrower tasks when limited datasets are available for classification, and the training and testing datasets are more consistent with each other. However, the correlation filters may have limitations [4], and their performance decreases when the intra-class variations are high, and the testing and training data are not consistent. The CNN-based feature extraction technique consist of a series of multi-scale non-linear filters, which increases its generalization capability as compared to a single trained correlation filter and provides a more robust strategy for gesture recognition, especially in challenging conditions. In our experiments, we have considered multiple backgrounds with fast and slow gesture movements as well as multi-gesture scenarios, which increase the intra-class variability in the training and testing datasets, thus degrading the performance of the non-linear correlation-based method proposed in [7].

The detection performance of the proposed approach has been compared using several performance metrics, which are summarized in Table 1. In addition to accuracy, we have also considered the F1 score and Mathew's correlation coefficient (MCC) for better evaluation of the performance of the proposed approach [38,39]. The F1 score represents the harmonic mean between precision and recall. Its value varies between 0 and 1, with 1 indicating perfect classification performance. The MCC indicates the correlation between the observed and predicted classification, and its value ranges between -1 to 1. An MCC value of 1 indicates a perfect classifier, and -1 indicates complete disagreement between true and predicted classes. The results have been summarized in Table 1.

From Table 1, we observe that in our experiments, the proposed system has a better performance in terms of all the metrics considered. In addition, the percentage reduction in AUC due to the effect of low illumination for various conditions considered in our experiments are significantly lower for the proposed approach as compared to the 2D imaging-based approach. In the case

Optics EXPRESS

of occlusion and multi-gesture scenarios, for all the metrics considered (as shown in Table 1) the proposed approach with low illumination performs better than even the 2D approach and other methodologies in high illumination as well as low illumination conditions which shows the effectiveness of the proposed approach as compared to other methodologies considered. Thus, the proposed 3D Integral imaging-based deep learning approach is promising for gesture recognition, especially under degraded environments.

4. Conclusion

In this paper, we have presented a system for human gesture classification using integral imaging and deep learning. Our experimental results show that using the proposed 3D integral imaging-based deep learning approach improves the performance under degraded environments such as occlusion and low light in comparison to the conventional 2D imaging approach and the other methodologies considered. In addition, the percent reduction in AUC for 3D imaging-based CNN-BiLSTM network due to degradation (simulated low illumination) is significantly lower than that of the 2D imaging-based approach. The proposed approach could be further extended for detecting human activities, sentiment analysis, etc., especially under more challenging conditions. Future work may consider other integral imaging approaches [40], comparison with time of flight sensing, and increased scene complexity.

Funding

Air Force Office of Scientific Research (FA9550-18-1-0338); Office of Naval Research (N000141712405, N000141712561).

Acknowledgments

T. O'Connor acknowledges support via the GAANN fellowship through the Department of Education. B. Javidi acknowledges support under Air Force Office of Scientific Research (FA9550-18-1-0338) and Office of Naval Research (N000141712405, N00014-17-1-2561).

Disclosures

The authors declare no conflicts of interest.

References

1. S. Mitra and T. Acharya, "Gesture Recognition: A Survey," IEEE Trans. Syst., Man, Cybern. C **37**(3), 311–324 (2007).
2. M. H. Kolekar and D. P. Dash, "Hidden Markov Model based human activity recognition using shape and optical flow based features," in *2016 IEEE Region 10 Conference (TENCON)* (2016), pp. 393–397.
3. I. Everts, J. C. van Gemert, and T. Gevers, "Evaluation of Color Spatio-Temporal Interest Points for Human Action Recognition," IEEE Trans. on Image Process. **23**(4), 1569–1580 (2014).
4. S. Ali and S. Lucey, "Are Correlation Filters Useful for Human Action Recognition?" in *2010 20th International Conference on Pattern Recognition* (2010), pp. 2608–2611.
5. S. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013).
6. Y. Zhang, Z. Zhang, Y. Zhang, J. Bao, Y. Zhang, and H. Deng, "Human Activity Recognition Based on Motion Sensor Using U-Net," IEEE Access **7**, 75213–75226 (2019).
7. X. Shen, H. Kim, K. Satoru, A. Markman, and B. Javidi, "Spatial-temporal human gesture recognition under degraded conditions using three-dimensional integral imaging," Opt. Express **26**(11), 13938–13951 (2018).
8. V. J. Traver, P. Latorre-Carmona, E. Salvador-Balaguer, F. Pla, and B. Javidi, "Three-Dimensional integral imaging for Gesture Recognition Under Occlusions," IEEE Signal Process. Lett. **24**(2), 171–175 (2017).
9. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature **521**(7553), 436–444 (2015).
10. S.-H. Hong, J.-S. Jang, and B. Javidi, "Three-dimensional volumetric object reconstruction using computational integral imaging," Opt. Express **12**(3), 483–491 (2004).
11. G. Lippmann, "Epreuves reversibles donnant la sensation du relief," J. Phys. **7**(1), 821–825 (1908).

Optics EXPRESS

12. C. B. Burckhardt, "Optimum Parameters and Resolution Limitation of Integral Photography," J. Opt. Soc. Am. **58**(1), 71–76 (1968).
13. N. Davies, M. McCormick, and L. Yang, "Three-dimensional imaging systems: a new development," Appl. Opt. **27**(21), 4520–4528 (1988).
14. F. Okano, H. Hoshino, J. Arai, and I. Yuyama, "Real-time pickup method for a three-dimensional image based on integral photography," Appl. Opt. **36**(7), 1598–1603 (1997).
15. M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, "Free-Viewpoint TV," IEEE Signal Process. Mag. **28**(1), 67–76 (2011).
16. M. Martínez-Corral and B. Javidi, "Fundamentals of 3D imaging and displays: a tutorial on integral imaging, light-field, and plenoptic systems," Adv. Opt. Photonics **10**(3), 512–566 (2018).
17. J.-S. Jang and B. Javidi, "Three-dimensional synthetic aperture integral imaging," Opt. Lett. **27**(13), 1144–1146 (2002).
18. B. Wilburn, N. Joshi, V. Vaish, E.-V. Talvala, E. Antunez, A. Barth, A. Adams, M. Horowitz, and M. Levoy, "High performance imaging using large camera arrays," ACM Trans. Graph. **24**(3), 765–776 (2005).
19. B. Javidi, R. Ponce-Díaz, and S.-H. Hong, "Three-dimensional recognition of occluded objects by using computational integral imaging," Opt. Lett. **31**(8), 1106–1108 (2006).
20. A. Stern, D. Aloni, and B. Javidi, "Experiments With Three-Dimensional Integral imaging Under Low Light Levels," IEEE Photonics J. **4**(4), 1188–1195 (2012).
21. A. Markman, X. Shen, and B. Javidi, "Three-dimensional object visualization and detection in low light illumination using integral imaging," Opt. Lett. **42**(16), 3068–3071 (2017).
22. I. Moon and B. Javidi, "Three-dimensional visualization of objects in scattering medium by use of computational integral imaging," Opt. Express **16**(17), 13080–13089 (2008).
23. M. Cho and B. Javidi, "Peplography—a passive 3D photon counting imaging through scattering media," Opt. Lett. **41**(22), 5401–5404 (2016).
24. R. Joshi, T. O'Connor, X. Shen, M. Wardlaw, and B. Javidi, "Optical 4D signal detection in turbid water by multi-dimensional integral imaging using spatially distributed and temporally encoded multiple light sources," Opt. Express **28**(7), 10477–10490 (2020).
25. H. Arimoto and B. Javidi, "Integral three-dimensional imaging with digital reconstruction," Opt. Lett. **26**(3), 157–159 (2001).
26. J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (2009), pp. 248–255.
27. T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust Object Recognition with Cortex-Like Mechanisms," IEEE Trans. Pattern Anal. Mach. Intell. **29**(3), 411–426 (2007).
28. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 1–9.
29. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Comput. **9**(8), 1735–1780 (1997).
30. J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), pp. 4694–4702.
31. M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," IEEE Trans. Signal Process. **45**(11), 2673–2681 (1997).
32. P. Baldi, S. Brunak, P. Frasconi, G. Soda, and G. Pollastri, "Exploiting the past and the future in protein secondary structure prediction," Bioinformatics **15**(11), 937–946 (1999).
33. X. Shen, A. Carnicer, and B. Javidi, "Three-dimensional polarimetric Integral imaging under low illumination conditions," Opt. Lett. **44**(13), 3230–3233 (2019).
34. B. Javidi and D. Painchaud, "Distortion-invariant pattern recognition with Fourier-plane nonlinear filters," Appl. Opt. **35**(2), 318–331 (1996).
35. P. Refregier, V. Laude, and B. Javidi, "Nonlinear joint-transform correlation: an optimal solution for adaptive image discrimination and input noise robustness," Opt. Lett. **19**(6), 405–407 (1994).
36. E. Keedwell, "An analysis of the area under the ROC curve and its use as a metric for comparing clinical scorecards," in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (2014), pp. 24–29.
37. S. Alam, O. Odejide, O. Olabiyi, and A. Annamalai, "Further results on area under the ROC curve of energy detectors over generalized fading channels," in *34th IEEE Sarnoff Symposium* (2011), pp. 1–6.
38. I. Martin-Diaz, D. Morinigo-Sotelo, O. Duque-Perez, and R. D. J. Romero-Troncoso, "Advances in Classifier Evaluation: Novel Insights for an Electric Data-Driven Motor Diagnosis," IEEE Access **4**, 7028–7038 (2016).
39. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics **21**(1), 6 (2020).
40. A. Castro, Y. Frauel, and B. Javidi, "Integral imaging with large depth of field using an asymmetric phase mask," Opt. Express **15**(16), 10266–10273 (2007).